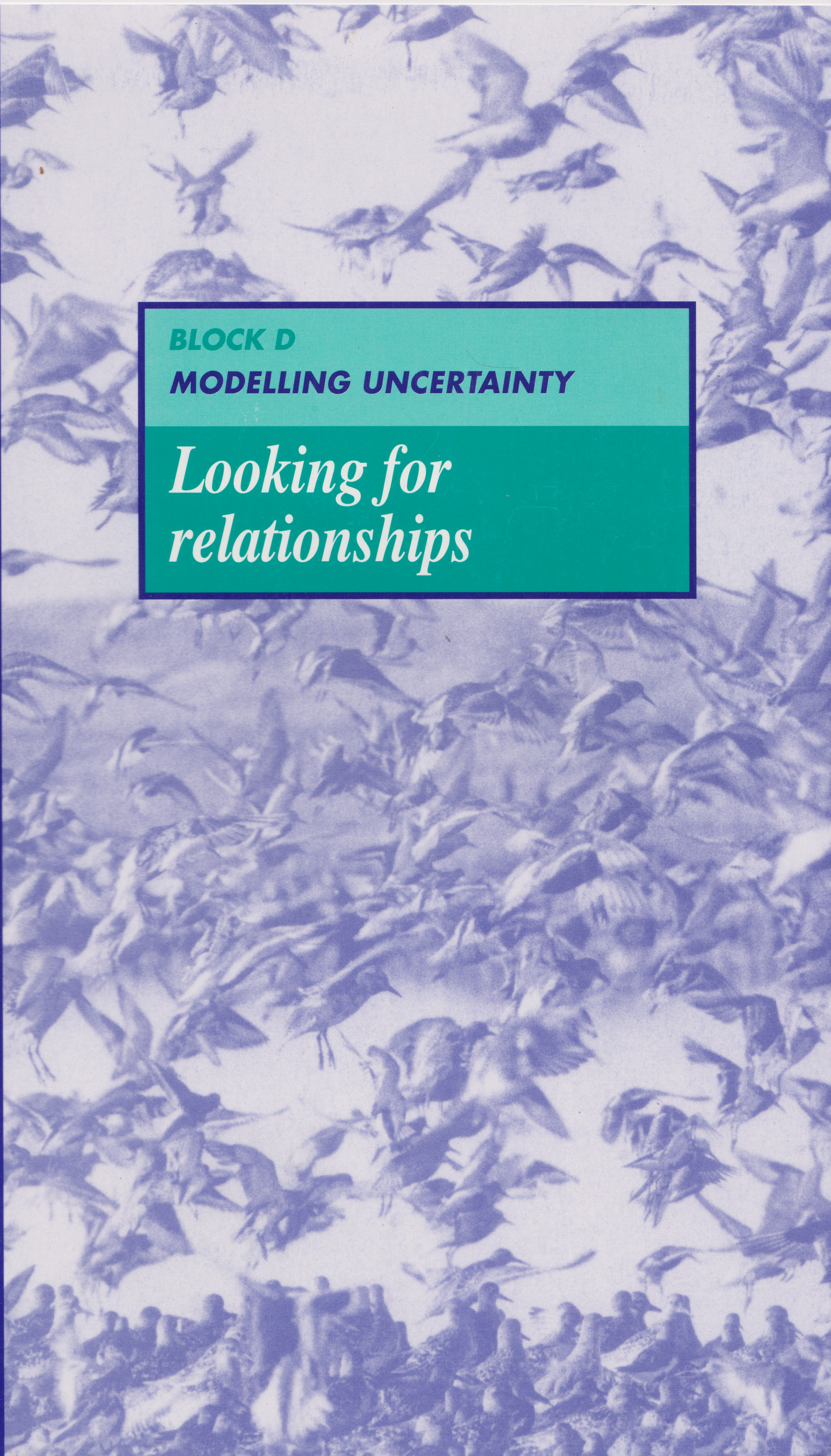
MST121 Chapter D5

The Open University

A first level interdisciplinary course

Using **Mathematics**



BLOCK D

MODELLING UNCERTAINTY

Looking for relationships

CHAPTER

D5

MST121 Chapter D5

The Open University

A first level interdisciplinary course

BLOCK D

MODELLING UNCERTAINTY

Looking for relationships

Using Mathematics

CHAPTER D5

Prepared by the course team

About this course

This course, MST121 *Using Mathematics*, and the courses MU120 *Open Mathematics* and MS221 *Exploring Mathematics* provide a flexible means of entry to university-level mathematics. Further details may be obtained from the address below.

MST121 uses the software program Mathcad (MathSoft, Inc.) and other software to investigate mathematical and statistical concepts and as a tool in problem solving. This software is provided as part of the course, and its use is covered in the associated Computer Book.

The Open University, Walton Hall, Milton Keynes MK7 6AA.

First published 1997. Reprinted 1997, 1998, 1999, 2000, 2001.

Edited, designed and typeset by The Open University using the Open University TEX System.

Printed in the United Kingdom by The Burlington Press, Foxton, Cambridge CB2 6SW.

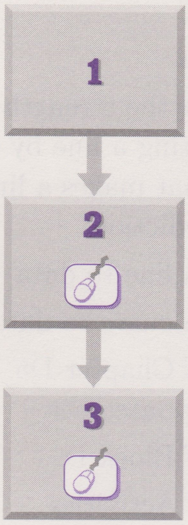
ISBN 0 7492 7811 0

This text forms part of an Open University First Level Course. If you would like a copy of *Studying with The Open University*, please write to the Course Enquiries Data Service, PO Box 625, Dane Road, Milton Keynes MK1 1TY. If you have not already enrolled on the Course and would like to buy this or other Open University material, please write to Open University Educational Enterprises Ltd, 12 Cofferidge Close, Stony Stratford, Milton Keynes MK11 1BY, United Kingdom.

1.6

Contents

Study guide



You should schedule three study sessions for your work on this chapter, of which the second and third will use the computer. The study pattern which we recommend is as follows.

Study session 1: Section 1.

Study session 2: Section 2. You will need access to your computer, together with the statistics software and Computer Book D.

Study session 3: Section 3. You will need access to your computer, together with the statistics software and Computer Book D.

None of these sessions should be long ones. If you chose to spend three sessions instead of two studying Chapter D4, then you may wish to study this chapter in two sessions. If so, then the study pattern we recommend is as follows.

Alternative study session 1: Sections 1 and 2 (computer).

Alternative study session 2: Section 3 (computer).

Alternatively, you could include all the computer work in one study session, by studying Section 1 in the first session and Sections 2 and 3 in the second. If you do this, then the second session may be a long one.

Introduction

In Chapter C3, functions were used to model a variety of phenomena: for instance, the velocity of a car moving with constant acceleration, population increase and radioactive decay. In each of these examples, the function described the relationship between two *variables*: velocity and time, population size and time, and the mass of a radioactive substance remaining and time. In another example, the chosen function modelled the relationship between the demand for a product and its price. And there are many examples of pairs of physical properties whose values depend on each other in a systematic way: for instance, altitude and atmospheric pressure, the volume of a metal object and its temperature, and the mean distance of a planet from the Sun and the time the planet takes to orbit the Sun.

The relationship between two variables may be established either empirically – that is, by consideration of data – or theoretically – that is, by reasoning from known laws. However, theories need to be verified by observation, so data are necessarily involved in either case. Once pairs of values of two variables have been obtained, a scatterplot of the data pairs can be drawn. We might hope that the points on the scatterplot will all lie exactly in a straight line, or exactly on a curve (that is, on a curve of simple shape). However, in many situations this will not be the case: even when it is evident from a scatterplot that there is a relationship between two variables, this relationship may not be an exact one.

Consider, for instance, the relationship between father's height and son's height, which you investigated in Section 4 of Chapter D2: a scatterplot of Pearson's data on the heights of 1078 father–son pairs showed that tall fathers tended to have tall sons, and short fathers short sons. However, the relationship between father's height and son's height is not exact – the heights of sons of 70-inch-tall fathers, for instance, are not all the same; they range from 64 inches to 78 inches. There is a lot of scatter in the plot.

In practice, physical measurements are subject to error and to the limitations of the equipment used, so, even when an exact relationship exists between two variables, there is likely to be some scatter in a plot of data. But if the points on a scatterplot do not lie exactly on a straight line or a curve, how do we decide which straight line or curve to choose to model the relationship?

In this chapter, we discuss the problem of choosing a line through a set of points. In Section 1, a criterion for choosing a line is discussed; this is the *principle of least squares*. Applying this principle leads to a method for choosing the line through a set of data points which, in one sense, is the 'best' line. This line is called the *least squares fit line* or the *regression line*. This method may be used whether or not the relationship sought is thought to be an exact one. In Section 2, you will be able to use OUStats to find the equation of the least squares fit line for Pearson's data, and also for several other data sets for which a linear function appears to be an appropriate model for the relationship between two variables.

In Section 3, we turn our attention briefly to modelling non-linear relationships. This section builds on Sections 1 and 2 and on some of the ideas introduced in Chapter C3.

Activity 0.1 Reviewing Block D

This short chapter is the last in Block D, so now would be a good time to review your progress on this block. One way in which you might begin to do this is by looking at the Handbook entries for Block D and using them to identify areas on which you need to do more work before the examination. You might find it useful to make three lists as follows. (You will not be able to complete these lists until after you have studied this chapter.)

- ◇ Identify ideas and techniques about which you are confident and which you will need to revise only briefly.
- ◇ Identify ideas and techniques which you are happy with but which you will nevertheless need to revise thoroughly.
- ◇ Identify anything about which you are unsure and on which you need to do more work.

You may find that carrying out this activity will help you to think about how you are going to set about reviewing the whole course and revising for the examination.

Another thing you might find it useful to do now is to note down anything in the assignment for Block D which you are unsure how to tackle. Is there anything that you do not understand and which you think you will be unable to sort out simply from the course materials?

We shall be returning to the theme of reviewing and revising Block D at the end of this chapter.

1 Checking the strength of concrete

Concrete is used widely in the construction industry. But not all concrete is the same. The concrete for a building (or a bridge, or whatever) must be specified at the design stage, since different constructions require concrete of different strengths. The crushing strength of concrete can be measured by subjecting cubes of concrete to increasing loads to determine when they will crumble. This means that samples of concrete can be tested in advance of construction. However, occasionally checks on the strength of the concrete in an existing structure become necessary. The concrete cannot be removed for testing, so how can checks be carried out?

Research has shown that the crushing strength of concrete is related to the speed with which an ultrasonic pulse passes through the concrete. So this speed, which is called the *pulse velocity*, can be used to predict the strength of concrete. But first the nature of the relationship between pulse velocity and crushing strength had to be established. Table 1.1 contains a typical set of data from an experiment to investigate this relationship. A scatterplot of the data is shown in Figure 1.1. The pulse velocity is plotted on the x-axis and crushing strength is plotted on the y-axis. (The crushing strength is the force per unit area required to crumble the concrete.)

It might have been more accurate to call this speed the 'pulse speed'; however, in practice, it is called the *pulse velocity*.

Table 1.1

Pulse velocity ($\mathrm{km\,s^{-1}}$) x	Crushing strength ($\mathrm{N\,mm^{-2}}$) y
3.91	15.1
3.94	17.2
4.00	12.2
4.07	14.9
4.24	25.1
4.25	21.0
4.32	23.0
4.39	25.0
4.40	25.0
4.41	26.0
4.42	29.0
4.43	23.5
4.44	29.0
4.50	30.4

$\mathrm{N\,mm^{-2}}$ is an abbreviation for 'newtons per square millimetre'.

These data are adapted from an example described on page 192 of *Probability and Statistics with Spreadsheets* by J. T. Callender and R. Jackson (Prentice Hall, 1995).

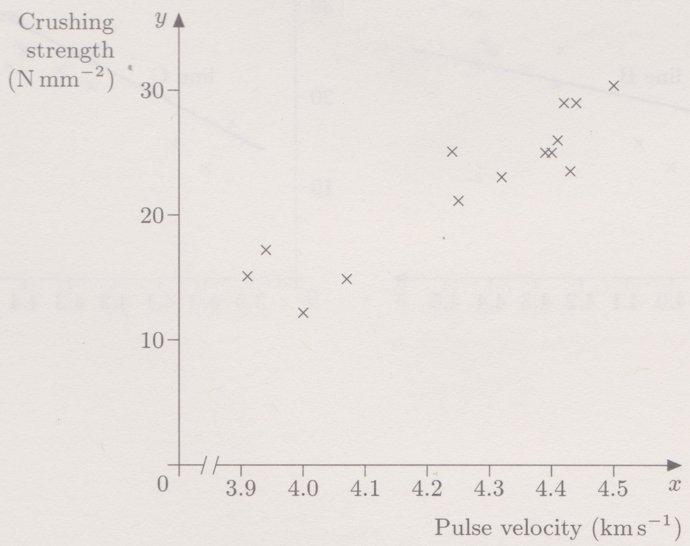


Figure 1.1 A scatterplot of crushing strength against pulse velocity

The scatterplot indicates that there is a relationship between pulse velocity and crushing strength, though there is some scatter in the plot. This could be the result of experimental error. Or it could be an indication that crushing strength and pulse velocity are not perfectly related, so that for any particular pulse velocity, a range of values of crushing strength is possible. Or the scatter may be due in part to both of these factors. However, it looks as though a straight line through the data might provide a useful summary of the relationship (at least for pulse velocities between 3.9 and 4.5 $\mathrm{km\,s^{-1}}$). But which line should we choose?

In this section, we discuss briefly two approaches to choosing a line to model the relationship between two variables. In Subsection 1.1, you will be introduced to an informal method – looking at a scatterplot of the data and choosing the line that in your opinion appears to fit the data best; this is called *fitting a line by eye*. Then, in Subsection 1.2, a formal method is described – a line is calculated using the data; this is called *the method of least squares*. In Subsection 1.3, we look briefly at how a chosen line may be used; for example, how can a line through the data in Figure 1.1 be used to predict the crushing strength of concrete for which the pulse velocity is 4.15 $\mathrm{km\,s^{-1}}$?

1.1 Fitting a line by eye

Choosing a line is often called **fitting** a line to the data. One method is simply to draw the line that you think best represents the relationship; this is known as **fitting by eye**. But it is not always clear, particularly when the points are widely scattered, which line to draw.

Activity 1.1 Which line?

Three attempts at fitting a line by eye to the concrete data are shown in Figure 1.2. In each diagram, x is the pulse velocity in $\mathrm{km\,s^{-1}}$ and y is the crushing strength in $\mathrm{N\,mm^{-2}}$. For each attempt, if you think the line could be improved, then say how you think this could be done.

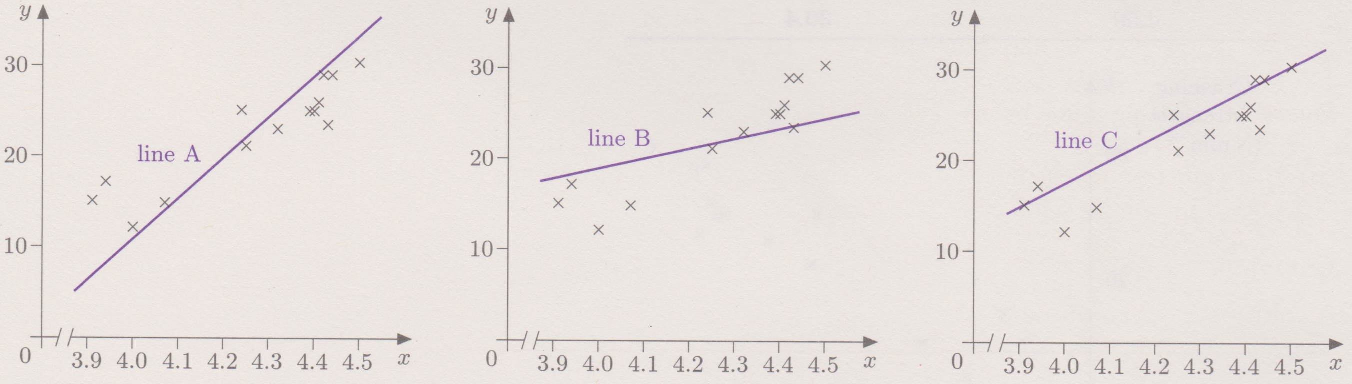


Figure 1.2 Three possible fit lines

Comment

If we used line A to predict the crushing strength of concrete, then we would underestimate the crushing strength when the pulse velocity is close to 3.9 and overestimate it when the pulse velocity is near 4.4 or 4.5: the line lies below all the points on the left-hand side of the diagram and above those on the right-hand side. So the line is too steep.

On the other hand, line B is not steep enough. In this case, the line lies above the points on the left-hand side of the diagram and below the points on the right-hand side.

Using line C, we would tend consistently to overestimate the strength of concrete: the line lies above nearly all of the points. It could be improved by lowering it a little so that it passes roughly through the 'middle' of the points.

Activity 1.2 Fitting a line by eye

(a) The above discussion suggests some factors to take into account when fitting a line by eye. Keeping these factors in mind, draw by eye on the scatterplot in Figure 1.3 the straight line which you think best summarises the relationship between pulse velocity and crushing strength.

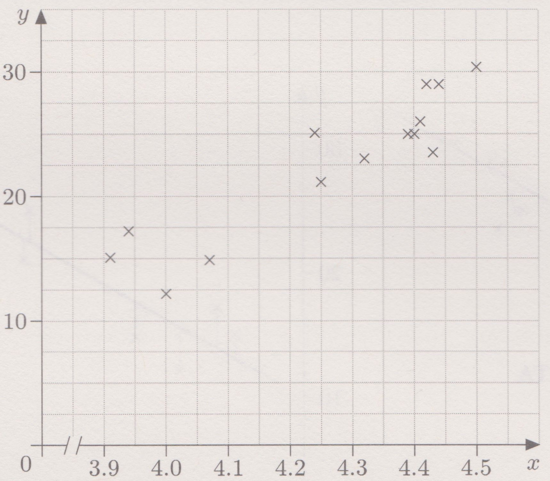


Figure 1.3 Fitting a line by eye

(b) Use your line to predict the crushing strength of concrete for which the ultrasonic pulse velocity is $4.15\,\mathrm{km\,s^{-1}}$.

Comment

A solution is given on page 23.

Your attempt at fitting a line by eye was probably slightly different from the one given in the solution. Fitting a line by eye is subjective, so no two people are likely to fit exactly the same straight line. There are many lines with similar slopes and at similar heights above the x-axis on the scatterplot which could be said to fit the data adequately. However, people using different lines would make different estimates for crushing strength. This is unsatisfactory if, for instance, decisions about the safety of a construction are to be based on such estimates. Instead of a subjective method, a well-defined procedure for choosing a 'good' line is needed. In the next subsection, we discuss what makes a line a 'good' fit, and describe an objective method for choosing a line.

In practice, the concrete in a construction is required to be considerably stronger than the minimum that theory suggests is necessary for the construction to be safe.

1.2 What makes a line a good fit?

In Activity 1.2, you fitted a line by eye to the concrete data and then used your line to predict the crushing strength of concrete for which the pulse velocity is $4.15\,\mathrm{km\,s^{-1}}$. One way of assessing how good a fit is provided by a line is to compare the recorded crushing strength of each sample for which we have data with the crushing strength predicted by the line; that is, by comparing the DATA (the recorded value) with the FIT (the predicted value). The DATA and FIT values are illustrated in Figure 1.4(a) for the data point $(4.24, 25.1)$ using the line a course team member fitted by eye (given in Solution 1.2). Also labelled on the diagram is the numerical difference between these values; this number is called the RESIDUAL.

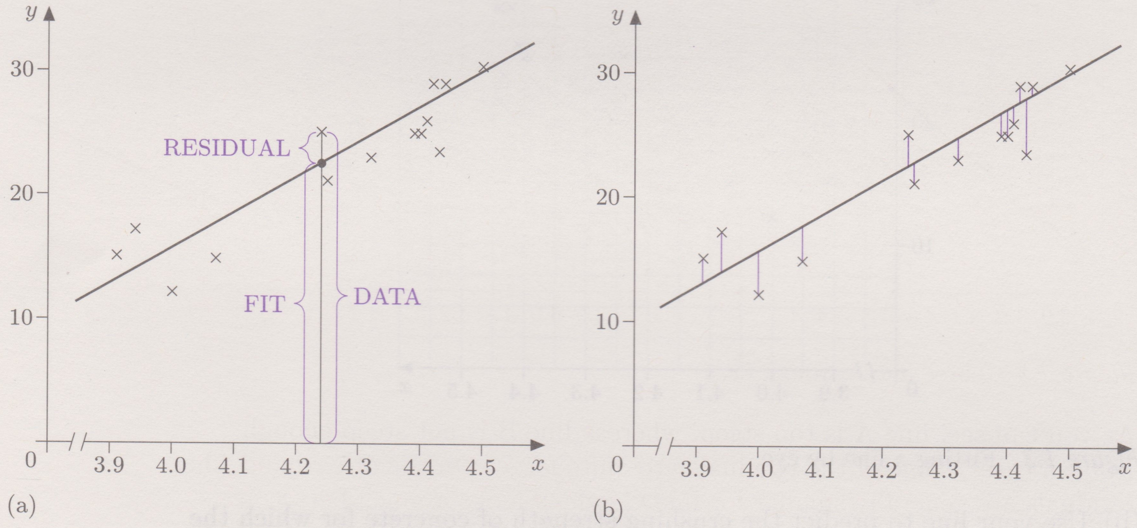


Figure 1.4 Residuals

The residuals are defined by the relationship

$$\text{RESIDUAL} = \text{DATA} - \text{FIT},$$

or, equivalently,

$$\text{DATA} = \text{FIT} + \text{RESIDUAL},$$

as is clear from Figure 1.4(a).

For a point above the fit line, the residual is positive; for a point below the fit line, the residual is negative. Intuitively, for a good fit line, we would like the residuals to be small and about half of them to be positive and half to be negative. For the line in Solution 1.2 fitted by eye, the residuals

are shown on Figure 1.4(b) for all the data points. As you can see, roughly half the residuals are positive and the rest are negative; and none of the residuals is very large. So the line appears to be a good fit.

Look again at the three lines in Figure 1.2. Line C is above most of the points, so most of the residuals are negative; clearly it is not a good fit. A line passing somewhere through the middle of the points, such as line A or line B, might be better. For each of these lines, about half of the points are above the line and half below. However, neither looks to be a good fit; line A is too steep and line B is not steep enough, so, for each line, some of the residuals are quite large. A line with intermediate slope would be better: most of the residuals would be smaller.

This suggests that a good fit line should have two properties: it should pass roughly through the 'middle' of the points, and its slope should be chosen so that the residuals are as small as possible. We try to select a line with these properties when we fit a line by eye but, as already noted, the choice using this method is subjective. What we need is an objective method of choosing a line with these properties.

One possibility is to choose a line passing through the point with coordinates $(\overline{x}, \overline{y})$, where $\overline{x}$ is the mean of the x-values and $\overline{y}$ is the mean of the y-values. Figure 1.5 shows three lines through the point $(\overline{x}, \overline{y}) = (4.266, 22.60)$ for the concrete data; the residuals are marked on each scatterplot.

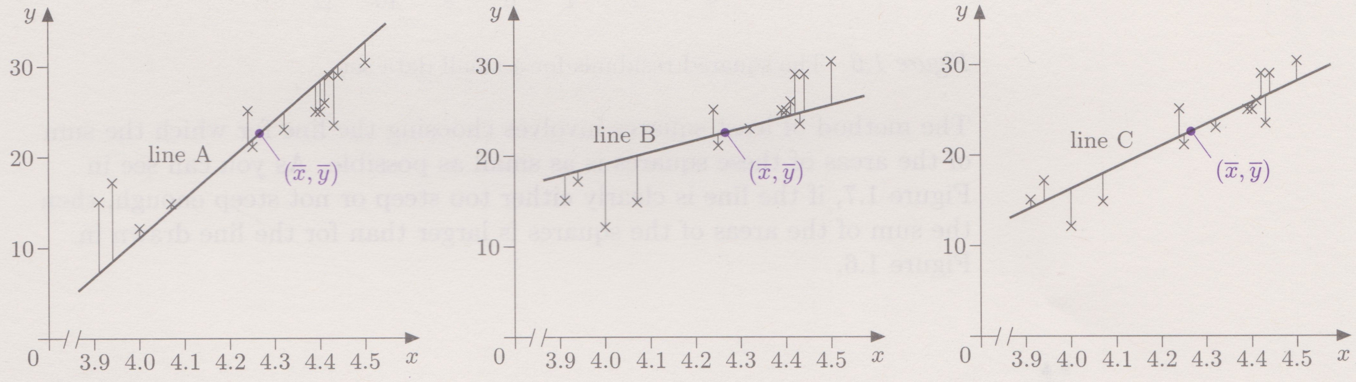


Figure 1.5 Three lines through $(\overline{x}, \overline{y})$

As you can see, line A is too steep, whereas line B is not steep enough. In both cases, some of the residuals are quite large. On the other hand, the residuals are generally smaller for line C, which looks a much better fit than either of the other two lines. We want the residuals to be as small as possible for the line that we choose. So why not choose the line for which the sum of the residuals is as small as possible? Unfortunately, this will not work: it can be shown that for any line through $(\overline{x}, \overline{y})$, the sum of the residuals is always zero – the sum of the positive residuals is always equal in magnitude to the sum of the negative residuals.

To overcome the problem of negative and positive residuals cancelling each other out, we could consider the sum of the magnitudes of the residuals and choose the line which makes *this* as small as possible. However, it is much more common to square the residuals – the squares are all positive or zero – and choose the line which minimises the sum of the squared residuals. This criterion for choosing a line is called the **principle of least squares**.

Applying this criterion leads to the most frequently used method of choosing a fit line: the **method of least squares**. The line chosen using this method is called the **least squares fit line** or the **regression line**.

One way of visualising the squared residuals is described below. For clarity, we shall use a small set of artificial data: Figure 1.6 shows four data points together with a possible fit line. For each point, the magnitude of the residual is equal to the vertical distance from the point to the line. The squared residual could be represented by the area of a square which has sides with length equal to the magnitude of the residual. In Figure 1.6, such a square is drawn for each point.

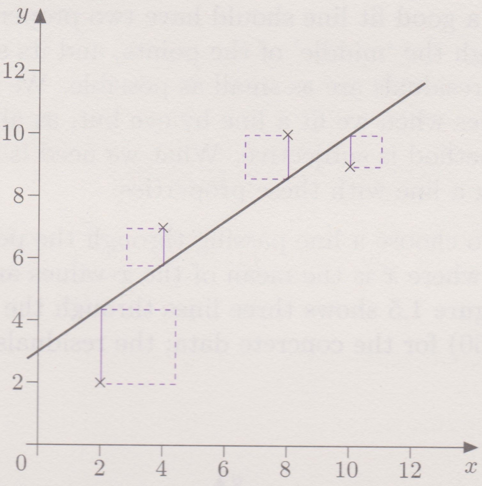


Figure 1.6 The squared residuals for a small data set

The method of least squares involves choosing the line for which the sum of the areas of these squares is as small as possible. As you can see in Figure 1.7, if the line is clearly either too steep or not steep enough, then the sum of the areas of the squares is larger than for the line drawn in Figure 1.6.

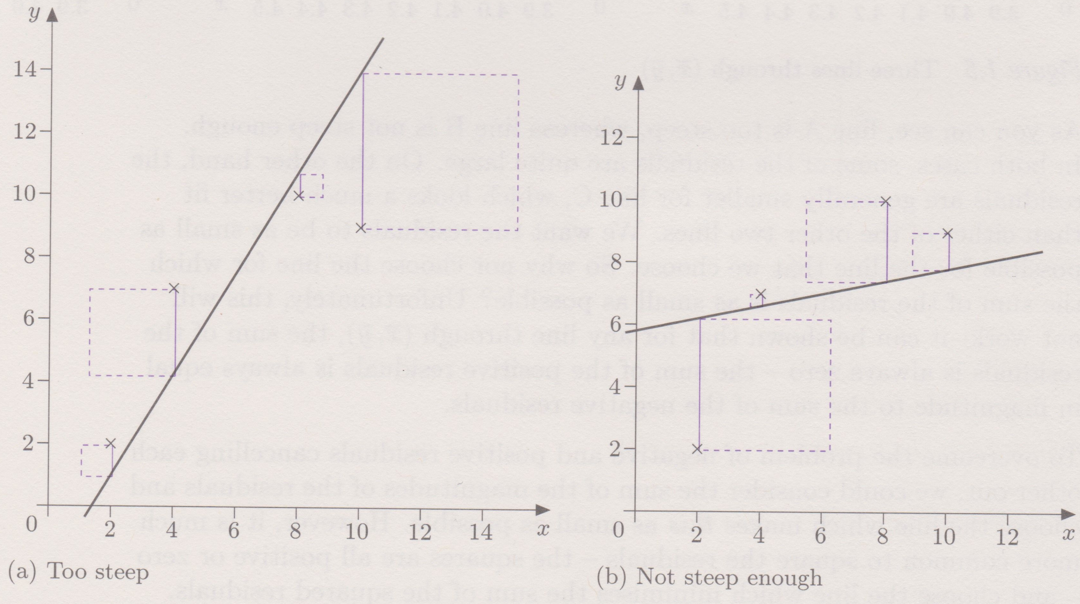
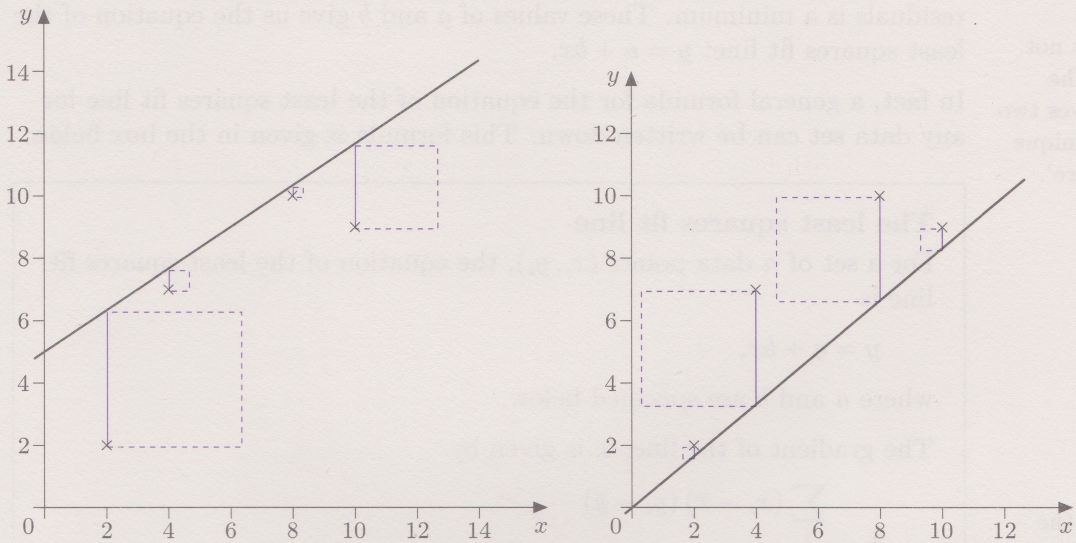


Figure 1.7 The squared residuals for two lines

Similarly, if the line is too high or too low, then the sum of the areas of the squares is also relatively large (see Figure 1.8).



(a) Too high (b) Too low

Figure 1.8 The squared residuals for two more lines

The least squares fit line is the line for which the sum of the squared residuals is as small as possible. The squared residuals are illustrated for the least squares fit line in Figure 1.9. In this example, you can see that the sum of the areas of the squares is smaller for this line than for any of the lines shown in Figures 1.6 to 1.8 (although it is not much smaller than for the line in Figure 1.6).

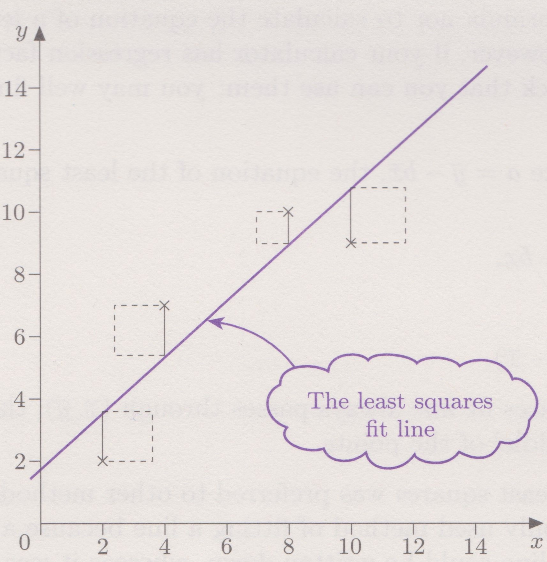


Figure 1.9 The squared residuals for the least squares fit line

It is possible to write down an algebraic expression for the sum of squared residuals which involves the gradient of the fit line (call it b) and the intercept of the line on the y-axis (call it a). Then, using either algebra or calculus, we can find the values of a and b for which the sum of squared residuals is a minimum. These values of a and b give us the equation of the least squares fit line: $y = a + bx$.

Using calculus involves the technique of partial differentiation (which is not covered in MST121). The algebraic method involves two applications of the technique of 'completing the square'.

In fact, a general formula for the equation of the least squares fit line for any data set can be written down. This formula is given in the box below.

The least squares fit line

For a set of n data points (x_i, y_i), the equation of the least squares fit line is

$$y = a + bx,$$

where a and b are specified below.

The gradient of the line, b, is given by

$$b = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

where $\overline{x}$, $\overline{y}$ are the means of the x-values and the y-values, respectively, and each sum is for i from 1 to n.

The intercept, a, is given by

$$a = \overline{y} - b\overline{x}.$$

If you are interested in the derivation of the formulas for a and b, then the details may be found in many statistics textbooks.

In this course, you will use your computer to calculate the equation of the least squares fit line for a set of data points. You will not be expected to remember this formula nor to calculate the equation of a least squares fit line by hand. However, if your calculator has regression facilities, then you may wish to check that you can use them: you may well find them useful in the future.

Notice that, since $a = \overline{y} - b\overline{x}$, the equation of the least squares fit line may be written as

$$y = \overline{y} - b\overline{x} + bx,$$

or

$$y - \overline{y} = b(x - \overline{x}).$$

So the least squares fit line always passes through $(\overline{x}, \overline{y})$; that is, it passes through the 'middle' of the points.

The method of least squares was preferred to other methods and became the most commonly used method of fitting a line because a formula for the least squares fit line could be written down, whereas it was not possible to write down a simple general formula for the fit line using other methods. For instance, there is no simple formula for the equation of the line which minimises the sum of the magnitudes of the residuals, and so finding the least squares fit line was preferred to finding this fit line.

We now return to the example of choosing a line to fit the concrete data. For these data, the equation of the least squares fit line is

$$y = -87.83 + 25.89x.$$

You will have the opportunity to verify this equation using OUStats in Section 2.

This line and the line fitted by eye (from Solution 1.2) are shown on the scatterplot in Figure 1.10. As you can see, the line fitted by eye is not the same as the least squares fit line. It is slightly steeper and passes close to but not through the point $(\overline{x}, \overline{y})$.

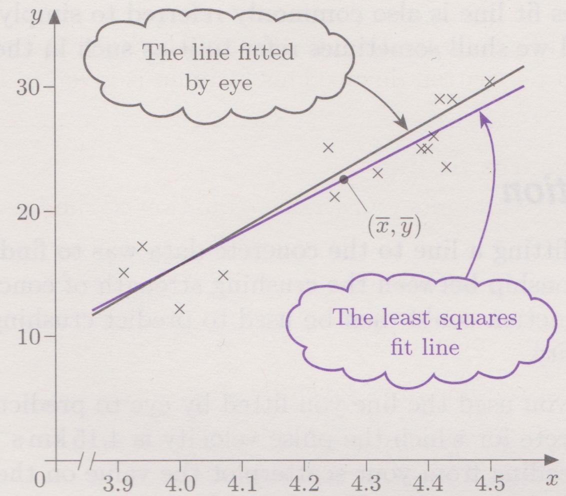


Figure 1.10 The least squares fit line and a line fitted by eye

Activity 1.3 Comparing the two fit lines

Add the least squares fit line to the scatterplot in Figure 1.3 on which you earlier drew a line by eye. (To draw the least squares fit line, you will need to find the coordinates of two points on it.)

How does the line you fitted by eye compare with the least squares fit line?

Comment

Some comments are given on page 23.

There is a great variety of terminology concerning model fitting in common usage in statistics. We have already mentioned that the least squares fit line is also called the regression line. More precisely, it is called **the regression line of y on x**. The variable x is called the **explanatory variable** (or sometimes the **independent variable**) and y is called the **dependent variable**. This language expresses the fact that we believe the value of x 'explains' in some degree the value of y: if we know the x-value, then we can predict the value of y. For the concrete data, we wanted to be able to predict the crushing strength of concrete from its pulse velocity, so crushing strength was the dependent variable and pulse velocity the explanatory variable.

Note that, when you wish to use a least squares fit line to predict values of one variable for values of another, the variable that you wish to predict is the dependent variable and must be plotted along the y-axis. The terminology 'the regression line of y on x' expresses the fact that this line is suitable for predicting y from x, and not vice versa. The line makes the *vertical* distances of the points from the line 'small'; to predict x from y you would want the horizontal distances from the points to the fit line to be 'small', so that the differences between the x-values and their predicted values are 'small'.

In fact, there is also a regression line of x on y, which may be used to predict x from y, and this is usually a different line from the regression line of y on x. The terminology thus distinguishes between these two lines. In this course, we shall be concerned only with the regression line of y on x, which is simply the least squares fit line.

The least squares fit line is also commonly referred to simply as the least squares line, and we shall sometimes refer to it as such in the rest of this chapter.

1.3 Prediction

Our purpose in fitting a line to the concrete data was to find a function to model the relationship between the crushing strength of concrete and pulse velocity. This function could then be used to predict crushing strength from pulse velocity.

In Activity 1.2, you used the line you fitted by eye to predict the crushing strength of concrete for which the pulse velocity is $4.15\,\mathrm{km\,s^{-1}}$. You did this simply by reading from your scatterplot the value on the fitted line of crushing strength (y) corresponding to a pulse velocity of $4.15\,\mathrm{km\,s^{-1}}$ ($x = 4.15$).

You could similarly use the least squares line to read from the scatterplot the value of crushing strength corresponding to a pulse velocity of $4.15\,\mathrm{km\,s^{-1}}$. Or you could use the equation of the least squares line to calculate this value – the FIT value. Since the equation of the least squares line is $y = -87.83 + 25.89x$, the FIT value corresponding to $x = 4.15$ is

$$y = -87.83 + 25.89 \times 4.15 \simeq 19.6.$$

So the least squares line predicts that the crushing strength of concrete for which the pulse velocity is $4.15\,\mathrm{km\,s^{-1}}$ is approximately $19.6\,\mathrm{N\,mm^{-2}}$. This is slightly smaller than the prediction of $20\,\mathrm{N\,mm^{-2}}$ arising from the line fitted by eye given in Solution 1.2. How does it compare with the prediction you made in Activity 1.2 using the line you fitted by eye?

Activity 1.4 Predicting the crushing strength

Use the equation of the least squares line to predict the crushing strength of concrete for which the pulse velocity is $4.3\,\mathrm{km\,s^{-1}}$.

Comment

A solution is given on page 23.

There are several observations that ought to be made at this stage. First, by predicting that the crushing strength is $19.6\,\mathrm{N\,mm^{-2}}$, we are not saying that the crushing strength of any concrete for which the pulse velocity is $4.15\,\mathrm{km\,s^{-1}}$ is *exactly* $19.6\,\mathrm{N\,mm^{-2}}$. The points on the scatterplot do not lie *exactly* on a straight line: they lie within a fairly narrow band on either side of the fit line. The predicted value is, in fact, an estimate of the *mean* crushing strength of concrete with pulse velocity $4.15\,\mathrm{km\,s^{-1}}$. The actual crushing strength of a particular sample of concrete may be higher or lower than the predicted value.

The precision of the predicted value can be indicated by giving a range of plausible values – a confidence interval – for the crushing strength of concrete with pulse velocity $4.15\,\mathrm{km\,s^{-1}}$ (or indeed for any other specified pulse velocity). Since the points lie fairly close to a straight line in this example, the confidence interval will be narrow. (For a scatterplot showing a lot of scatter, such a confidence interval would be wide.)

If you study statistics further in the future, then you may well learn how to calculate such confidence intervals. However, we shall not discuss how they are calculated in MST121. The point to remember is that the predicted value is only an estimate of the strength of the concrete. If safety is an issue, then we would want to know how precise the estimate is, so we would need a confidence interval as well as the predicted value.

Such confidence intervals are discussed in the course M246, for instance.

The second observation that we should make is that predictions from a fit line are valid only for the range of values of x, the explanatory variable, that are represented in the data. For the concrete data, the values of pulse velocity in the data range from 3.91 to 4.50. It is possible that the straight-line model would not be appropriate for values outside this range, so we should not use the fit line to predict crushing strength for pulse velocities such as 3.7 or 4.7. It is reasonable to use the fit line to make predictions for pulse velocities only within, or just outside, the range of values in the data.

The final observation is that, as already mentioned at the end of Subsection 1.2, the least squares line may be used to predict crushing strength (y) for values of pulse velocity (x), but not vice versa.

Summary of Section 1

In this section, two methods of choosing a line to model the relationship between two variables have been discussed: *fitting a line by eye* and *calculating the least squares fit line.*

When a line is fitted to a set of data points, the residual of each data pair may be calculated using the relationship

$$\text{RESIDUAL} = \text{DATA} - \text{FIT},$$

where DATA is the y-coordinate of the data pair, and FIT is the y-value predicted by the line for the corresponding x-coordinate.

The least squares fit line is the line which minimises the sum of the squared residuals for the data set. It is often referred to simply as the *least squares line* and is also known as the *regression line of y on x*. It may be used to predict values of y, the *dependent variable*, for values of x, the *explanatory variable*, but not vice versa. It should be used only to predict y-values for x-values within, or just outside, the range of values of x represented in the data.

2 Fitting a line to data

When choosing a function to model the relationship between two variables, the first step is to obtain a scatterplot of the data. If a straight-line model seems appropriate, then a line can be fitted either by eye or using the method of least squares. In this section, you will be using OUStats to investigate several data sets containing paired data.

Refer to Computer Book D for the work in this section.

Summary of Section 2

In this section, you have used OUStats to explore the relationship between several pairs of variables. You have revisited Pearson's data on the heights of fathers and sons (from Chapter D2), and the data on memory and age (from Chapter D4), and you have investigated the pattern of eruptions of the Old Faithful geyser (using data collected in August 1978). The use of OUStats to calculate the equation of the regression line of y on x and to draw the regression line on a scatterplot has been described.

3 Fitting a curve to data

The method of least squares can also be used when a scatterplot of data suggests that a suitable model for the relationship between two variables is a curve rather than a straight line. In Chapter C3, you saw that when the relationship between two variables is not linear, it is sometimes possible to transform the data so that the relationship between the transformed variables may be modelled by a straight line. The relationship between the original variables can then be deduced.

There are many possible ways of transforming data: squaring, cubing, taking the square root, taking logarithms, etc. Different data sets require different transformations to produce a plot with a roughly linear pattern. However, in this section we shall restrict our attention to transformations using logarithms. In Chapter C3, a log–log plot was used where it was thought that two variables were related by a power law, and a log–lin plot was used where an exponential function was thought to be an appropriate model for a relationship. In this section, we shall build on these ideas in order to fit a curve to data in several examples where the relationship between the variables is clearly not a linear one.

Essentially, the method we shall use is to transform the data and then fit a straight line to the transformed data using the method of least squares. This line corresponds to a curve through the original data, and thus we can deduce the equation of a curve which models the relationship between the original variables. We shall apply this method only to data sets which may be successfully transformed using logarithms. We shall not discuss how to select a suitable transformation when logarithms fail.

Other transformations are discussed briefly in the course M246.

All the calculations in this section will be carried out using OUStats. The method will be illustrated for some data on the population of the USA.

Refer to Computer Book D for the work in this section.

Summary of Section 3

In this section, you have explored three data sets. In each case, a scatterplot of the data indicated that the two variables in the data set are related, but that a straight line is not a suitable model for the relationship. A curve was fitted to the data using the following procedure.

◇ First the data for either one or both of the variables are transformed using logarithms, and a scatterplot of the transformed data is obtained.

◇ If the scatterplot indicates that a straight line is an appropriate model for the relationship between the transformed variables, then the method of least squares is used to fit a line to the transformed data.

◇ The equation of the corresponding curve through the original data is then written down. This curve is a model for the relationship between the original variables.

The use of OUStats to transform data was explained.

Summary of Chapter D5

This chapter has been concerned with exploring the relationship between two variables and choosing a line or curve to model the relationship. You have been introduced to the most commonly used method of fitting a line to data – the method of least squares. For several data sets, you used OUStats to choose a line to model the relationship between two variables. The use of the least squares line for prediction was discussed.

Log–lin plots and log–log plots were reviewed briefly. You saw how least squares may be used to fit a line, where it looks as though the pattern in a log–lin plot or a log–log plot of the data could be modelled by a straight line, and how fitting a line to one of these plots is equivalent to fitting a curve to the original data.

Learning outcomes

You have been working towards the following learning outcomes.

Terms to know and use

Dependent variable, explanatory variable, the least squares fit line, the regression line of y on x, residual, the predicted or FIT value.

Ideas to be aware of

◇ How residuals can be used to judge whether a line is a good fit to a set of data.

◇ That the least squares fit line is the straight line for which the sum of the squared residuals is a minimum.

◇ How the least squares line is used to predict the values of the dependent variable for values of the explanatory variable.

◇ That the least squares line should be used for prediction only for values of the explanatory variable within, or possibly just outside, the range of values included in the data.

Features of OUStats to use

◇ Include the least squares fit line on a scatterplot.

◇ Obtain the equation of the least squares fit line.

◇ Transform data using logarithms.

Summary of Block D

This block has been concerned principally with statistical ideas. You have been introduced to probability as a way of modelling the uncertainty inherent in a situation, and to probability distributions as models for variation.

The links between populations and samples have been discussed; in particular, you have been introduced to sampling distributions and to the central limit theorem.

Much of statistics is concerned with using samples of data to infer information about the populations from which the samples were drawn. The last three chapters have been concerned with how this can be done. Each chapter involved a different type of investigation and contained a brief introduction to an important statistical idea. In Chapter D3, you met confidence intervals; in Chapter D4, hypothesis testing was discussed; and in Chapter D5, you were introduced to fitting models to data using regression. These ideas underpin a large body of statistical techniques.

You have also used specially-designed software to explore problems involving chance and to consolidate your understanding of the nature of confidence intervals. And you have used the data analysis package OUStats to explore and analyse a variety of data sets.

We hope that this introduction to probability and statistics will enable you to approach any statistical aspects of your future studies with confidence.

At the beginning of this chapter, you began the process of reviewing Block D. You are invited to continue this review and extend it to the whole course in the two activities below. You are also asked to consider how you are going to set about planning your revision for the examination.

Activity 4.1 Reviewing Block D

Look back at the notes you made on Activity 0.1. Now that you have studied this chapter, do you wish to amend the lists you made? If so, then do it now. And if you identified anything in the assignment that you were unsure how to tackle, do you now understand what to do? If not, then have you sought advice from your tutor yet?

Activity 4.2 Reviewing and revising MST121

Reviewing your progress on Block D should have given you some ideas about how you might tackle the task of revising for the examination. If your revision is to be effective, then it needs to be organised – you need to make a plan. A good starting point is to review the course. In Activity 0.1, we suggested how you might use the Handbook entries for Block D to classify the topics in this block according to how much further work you need to do on them. You might find it useful to carry out a similar exercise for the other blocks of the course. You can then use your notes to decide how much time to allocate to the different parts of the course as you revise for the examination.

Other points worth considering at this stage include the following.

- ◇ When are you going to tackle the specimen examination paper, and how are you going to use it?
- ◇ What sources of help are available to you during the revision period?

Note down your thoughts before reading the comments below.

Comment

It is probably not a good idea to use the specimen examination paper to draw up your plan for revision. The questions will not be exactly the same every year, so if you revise only what you need in order to do this paper, then you may fail to revise other important topics. However, when you have made a list of topics that you plan to revise, you may like to check that it includes everything you need for the specimen paper.

You may find it useful to try the specimen paper as a trial run for the examination. In this case, it would be a good idea to leave it until you have completed your basic revision, but while you still have time to look again at anything on the paper that causes you difficulty. Try doing the specimen paper using only the materials you will have with you in the examination – the Handbook and a calculator. This may suggest to you some further useful annotation for the Handbook.

There are many sources of help available to you between now and the examination. The course materials, your assignments with comments from your tutor, and the specimen examination paper will all be useful. And if you cannot sort out something for yourself or are unsure of some idea, then talk with others about it – with fellow students, or with your tutor, or using the telephone helpline.

Finally, do not postpone making your revision plan for too long. If you need help with this, then do seek advice from your tutor.

Solutions to Activities

Solution 1.2

(a) A course team member's attempt at fitting a line by eye is shown in Figure S.1.

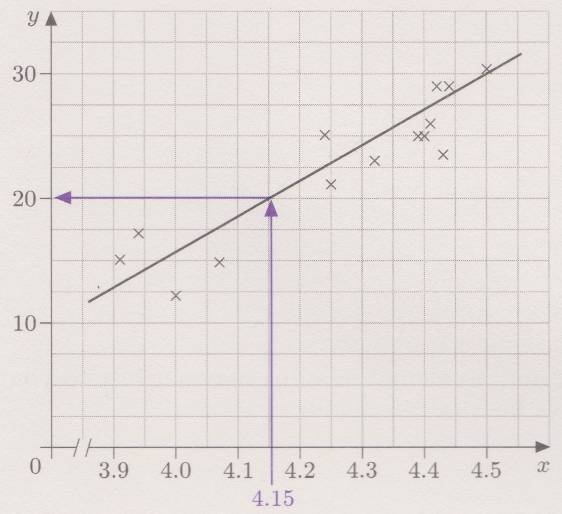


Figure S.1 A line fitted by eye

(b) Using the fitted line, the estimate is $20.0\,\mathrm{N\,mm^{-2}}$ for the crushing strength of concrete for which the ultrasonic pulse velocity is $4.15\,\mathrm{km\,s^{-1}}$.

Solution 1.3

The equation of the least squares fit line is $y = -87.83 + 25.89x$. When $x = 3.9$, for instance,

$$y = -87.83 + 25.89 \times 3.9 \simeq 13.1,$$

and when $x = 4.5$,

$$y = -87.83 + 25.89 \times 4.5 \simeq 28.7,$$

so the line passes through the points $(3.9, 13.1)$ and $(4.5, 28.7)$. Notice that the coordinates found were of two points whose x-coordinates were at either end of the range of values covered by the scatterplot. This is good practice, as it is easier to draw a line accurately if your two points for determining the line are a long way apart than if they are close together.

Solution 1.4

When $x = 4.3$,

$$y = -87.83 + 25.89 \times 4.3 \simeq 23.5.$$

So the predicted crushing strength of concrete for which the pulse velocity is $4.3\,\mathrm{km\,s^{-1}}$ is approximately $23.5\,\mathrm{N\,mm^{-2}}$.

Using Mathematics

The Open University
ISBN 0 7492 7811 0